AEM: Problem Set 1

Maxwell Austensen

September 24, 2016

```
# Utility functions ------------------------------------------------------

`%S%` <- function(x, y) {
  paste0(x, y)
}

`%notin%` <- Negate(`%in%`)

##############################################################################

# Install packages if needed
package_list <- c("stargazer", "knitr", "haven", "labelled", "ICC", "scales", "tidyverse")
new_packages <- package_list[package_list %notin% installed.packages()[,"Package"]]
if(length(new_packages)) install.packages(new_packages)

library(stargazer)
library(knitr)
library(haven)
library(labelled)
library(ICC)
library(scales)
library(tidyverse)

# Set directories
repo_ <- "H:/GitHub/aem/"
ps1_ <- "C:/Users/austensen/Box Sync/aem/ps1/"

##############################################################################
```

```
data_raw <-
  read_stata(ps1_ %S% "Thornton HIV Testing Data.dta") %>%
  remove_val_labels

names(data_raw) <- names(data_raw) %>% tolower

main_sample <-
  data_raw %>%
  filter(
    hiv2004 %notin% c(NA, -1),
    !is.na(any),
    !is.na(zone),
    !is.na(age)
  )
```

Part I: Summary Statistics

```
get_summary <- function(data){
  data %>%
    summarise(
      `Average Age` = mean(age, na.rm=T),
      `Percentage of Males` = mean(male, na.rm=T)*100,
      `Average Years of Education` = mean(educ2004, na.rm=T),
      `Percentage with HIV` = mean(hiv2004, na.rm=T)*100
    ) %>%
    kable(digits = 1)
}
```

1.

```
main_sample %>% get_summary
```

Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
33.4	46.3	3.6	6.3

2.

```
main_sample %>% group_by(any) %>% get_summary
```

any	Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
0	32.1	47.1	4.5	6.3
1	33.7	46.1	3.4	6.2

```
main_sample %>% group_by(under) %>% get_summary
```

under	Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
0	33.7	46.5	3.5	5
1	33.0	46.0	3.9	8

There are no major differences in the variables betweeen treatment and control groups based on either cash recipt or distance. However, those that received some cash were on average about a year old and 2% more likely to be male than the control group.

3.

```
grps <- c("any", "under")
vars <- c("age", "male", "hiv2004")

ttest_results <-
  data.frame(
    var = character(),
    P.value = double(),
    stringsAsFactors=FALSE
  )

for(grp in grps) {
  i <- 1
  for(var in vars){
    result <- t.test(main_sample[[var]] ~ main_sample[[grp]], var.equal = TRUE)$p.value

    ttest_results[i,1] <- var
    ttest_results[i,2] <- result
    i <- i+1
  }
  writeLines("#### t-test: group = " %S% grp)
  print(kable(ttest_results, digits = 3))
  writeLines("\n\n\n")
}
```

t-test: group = any

var	P.value
age	0.008
male	0.658
hiv2004	0.952

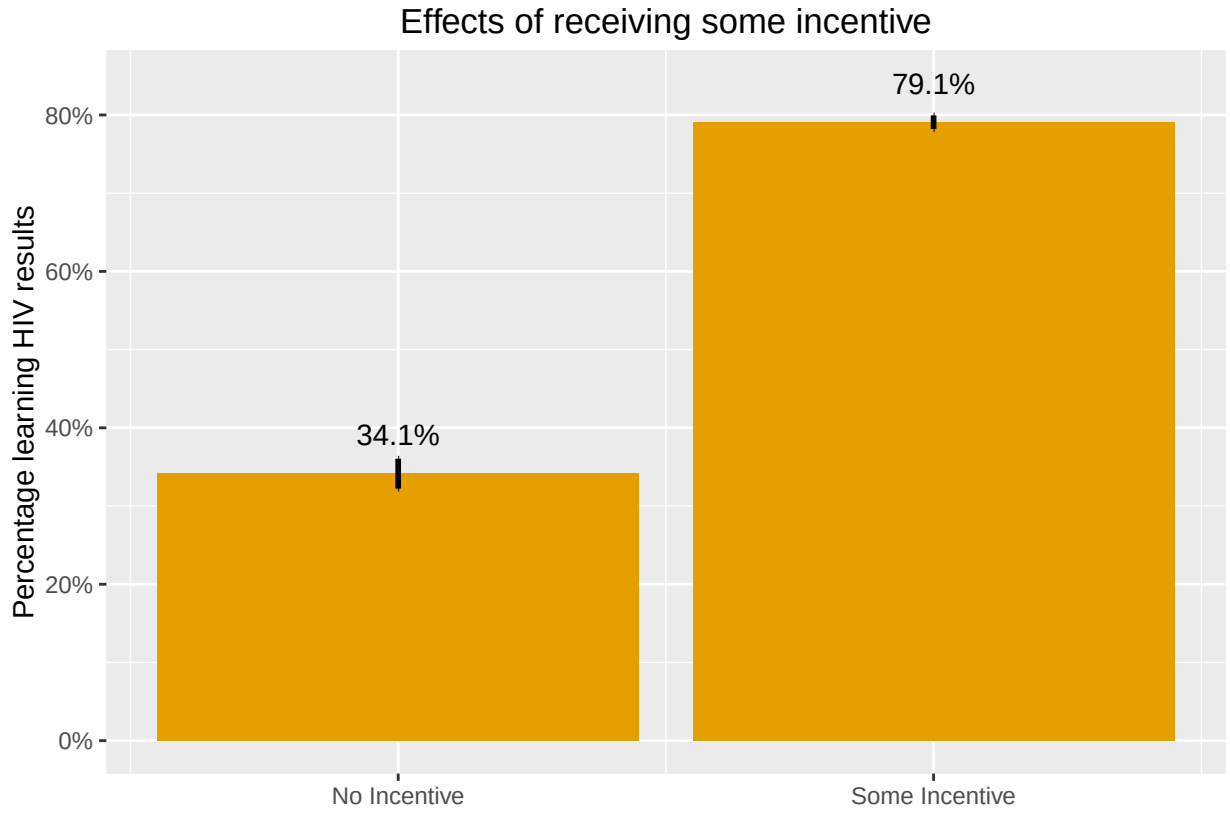
t-test: group = under

var	P.value
age	0.194
male	0.813
hiv2004	0.001

Part II: Analysis using graphs

4.

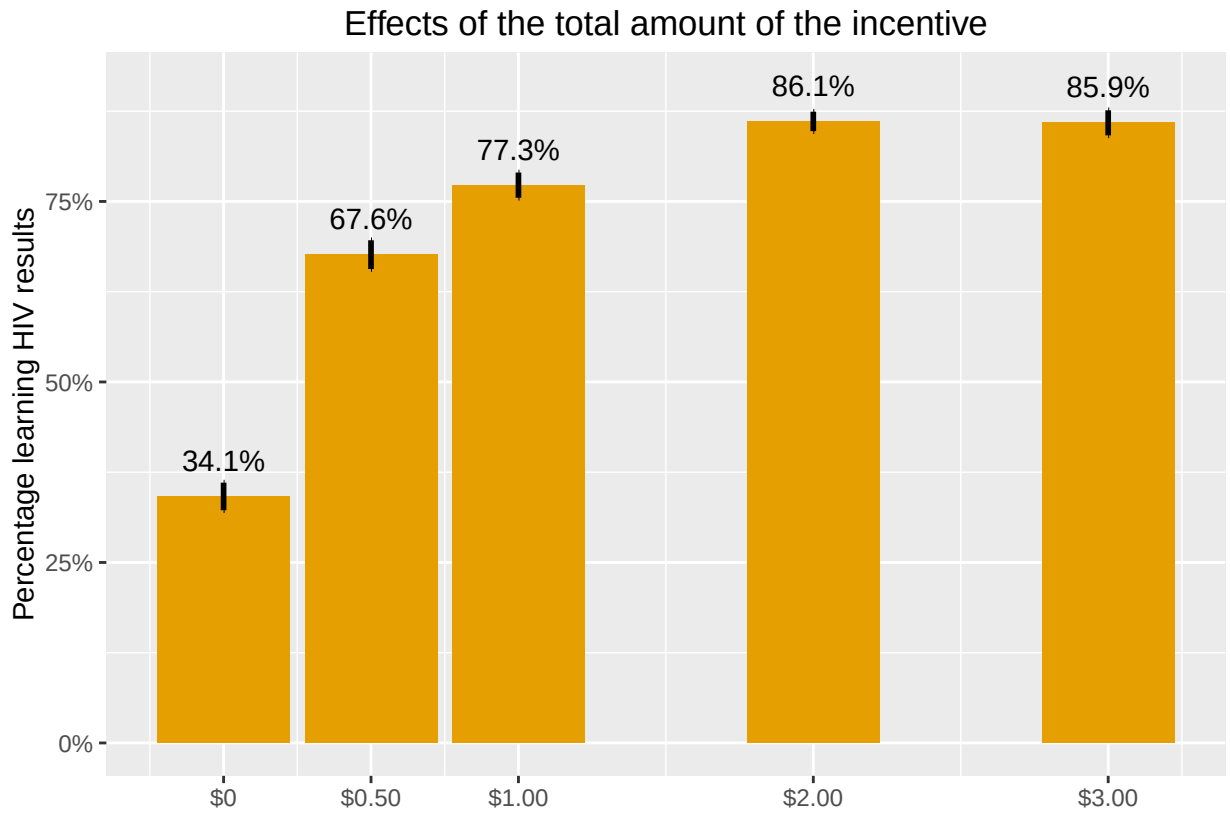
```r
main_sample %>%
  group_by(any) %>%
  summarise(
    got_mean = mean(got, na.rm=T),
    got_se = sd(got)/sqrt(n())
  ) %>%
  ggplot(aes(any, got_mean)) +
  geom_bar(stat = "identity", fill = "#E69F00") +
  geom_errorbar(aes(ymin = got_mean - got_se, ymax = got_mean + got_se), width = 0, size = 1) +
  geom_text(aes(y = got_mean+0.05, label = round(got_mean*100, 1) %S% "%")) +
  scale_y_continuous(labels = percent_format()) +
  scale_x_continuous(breaks = c(0,1), labels = c("No Incentive", "Some Incentive")) +
  ylab("Percentage learning HIV results") + xlab("") +
  ggtitle("Effects of receiving some incentive")
```



5.

```r
main_sample %>%
  group_by(ti) %>%
  summarise(
```

```
  got_mean = mean(got, na.rm=T),
  got_se = sd(got)/sqrt(n())
) %>%
ggplot(aes(ti, got_mean)) +
geom_bar(stat = "identity", fill = "#E69F00") +
geom_errorbar(aes(ymin = got_mean - got_se, ymax = got_mean + got_se), width = 0, size = 1) +
geom_text(aes(y = got_mean+0.05, label = round(got_mean*100, 1) %S% "%")) +
scale_y_continuous(labels = percent_format()) +
scale_x_continuous(breaks = c(0,50, 100, 200, 300), labels = c("$0", "$0.50", "$1.00", "$2.00", "$3.00
ylab("Percentage learning HIV results") + xlab("") +
ggtitle("Effects of the total amount of the incentive")
```



Part III: Analysis using linear regression

6.

```
any_1 <- lm(got ~ any, data = main_sample)
any_2 <- lm(got ~ any + age + male + educ2004 + mar, data = main_sample)

stargazer(any_1, any_2, type = output_type)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Sep 25, 2016 - 10:58:22 PM

Table 6:

	Dependent variable:	
	got	
	(1)	(2)
any	0.449***	0.450***
	(0.019)	(0.020)
age		0.001
		(0.001)
male		−0.010
		(0.017)
educ2004		−0.009***
		(0.003)
mar		0.013
		(0.022)
Constant	0.341***	0.362***
	(0.017)	(0.039)
Observations	2,812	2,530
R^2	0.162	0.181
Adjusted R^2	0.162	0.179
Residual Std. Error	0.423 (df = 2810)	0.412 (df = 2524)
F Statistic	545.100*** (df = 1; 2810)	111.239*** (df = 5; 2524)
Note:		*p<0.1; **p<0.05; ***p<0.01

7.

```
group_diff_any <-
  main_sample %>%
  group_by(any) %>%
  summarise(got = mean(got*100, na.rm=T)) %>%
  spread(any, got) %>%
  transmute(diff = `1` - `0`) %>%
  .[[1]]
```

Using a group means comparison, the estimated treatment effect is 44.94. This answer does not differ significantly from the OLS coefficent estimate on treatment. Since the addition of control variables does not significatly alter the treatment effect estimate, this suggests that the randomization of treatment was successful in balancing the two groups with respect to these other variables.

8.

```
ti_1 <- lm(got ~ ti, data = main_sample)
ti_2 <- lm(got ~ ti + age + male + educ2004 + mar, data = main_sample)

stargazer(ti_1, ti_2, type = output_type)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Sep 25, 2016 - 10:58:22 PM

Table 7:

	Dependent variable:	
	got	
	(1)	(2)
ti	0.002***	0.002***
	(0.0001)	(0.0001)
age		0.001
		(0.001)
male		−0.024
		(0.018)
educ2004		−0.013***
		(0.003)
mar		0.003
		(0.022)
Constant	0.499***	0.551***
	(0.013)	(0.037)
Observations	2,812	2,530
R^2	0.125	0.142
Adjusted R^2	0.125	0.141
Residual Std. Error	0.432 (df = 2810)	0.422 (df = 2524)
F Statistic	401.536*** (df = 1; 2810)	83.879*** (df = 5; 2524)
Note:		*p<0.1; **p<0.05; ***p<0.01

9.

Part IV: Conditional (Heterogeneous) Treatment Effects

10.

```
any_male <- lm(got ~ any + male + any*male, data = main_sample)

stargazer(any_male, type = output_type)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Sep 25, 2016 - 10:58:22 PM

Table 8:

	Dependent variable:
	got
any	0.445^{***}
	(0.026)
male	-0.015
	(0.034)
any:male	0.009
	(0.039)
Constant	0.349^{***}
	(0.023)
Observations	2,812
R^2	0.163
Adjusted R^2	0.162
Residual Std. Error	0.423 (df = 2808)
F Statistic	181.701^{***} (df = 3; 2808)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimate for the treatment-male interaction term is -0.015. It is not statistically significant. This suggests that there is not differential effect of recieving any cash ammount for men and women.

11.

```
any_educ <- lm(got ~ any + educ2004 + any*educ2004, data = main_sample)

stargazer(any_educ, type = output_type)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Sep 25, 2016 - 10:58:22 PM

Table 9:

	Dependent variable:
	got
any	0.446^{***}
	(0.030)
educ2004	-0.010^{**}
	(0.005)
any:educ2004	0.001
	(0.005)
Constant	0.394^{***}
	(0.027)
Observations	2,530
R^2	0.180
Adjusted R^2	0.179
Residual Std. Error	0.412 (df = 2526)
F Statistic	184.730^{***} (df = 3; 2526)
Note:	*p<0.1; **p<0.05; ***p<0.01

Part V: Policy Implications

12.

13.

Part VI: A Random Sub-Sample

14.

```
sample_1000 <- main_sample %>% sample_n(1000)
```

15.

```
any_1 <- lm(got ~ any, data = sample_1000)
any_2 <- lm(got ~ any + age + male + educ2004 + mar, data = sample_1000)

stargazer(any_1, any_2, type = output_type)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Sep 25, 2016 - 10:58:22 PM

Table 10:

	Dependent variable:	
	got	
	(1)	(2)
any	0.444***	0.444***
	(0.033)	(0.034)
age		0.0003
		(0.001)
male		0.011
		(0.030)
educ2004		−0.008*
		(0.005)
mar		0.028
		(0.038)
Constant	0.336***	0.347***
	(0.029)	(0.067)
Observations	1,000	902
R^2	0.157	0.177
Adjusted R^2	0.156	0.172
Residual Std. Error	0.428 (df = 998)	0.417 (df = 896)
F Statistic	185.802*** (df = 1; 998)	38.466*** (df = 5; 896)
Note:		*p<0.1; **p<0.05; ***p<0.01

Part VII: Choosing Sample Size

16.

17.

```
main_sample %>%
  filter(!is.na(usecondom04)) %>%
  ICCbare(site, usecondom04, data = .)
```

```
## [1] 0.01457493
```

Part VIII: Fisher Randomization Test (bonus)

18.

```
# data %>%
#   filter(!is.na(any)) %>%
#   group_by(any) %>%
#   summarise(got = mean(got, na.rm=T))
#
# oneway_test(got ~ factor(any), data, alternative = "greater")
# independence_test(got ~ factor(any), data,
#                   alternative = "greater",
#                   distribution = approximate(B = 1000))
#
# fisher.test(data$got, factor(data$any), alternative = "greater", simulate.p.value = TRUE)
```
